

# Realizacja hurtowni danych dla administracji publicznej na przykładzie budowy systemu IACS

*Mariusz Muszyński*

Pentacomp Systemy Informatyczne

Prace nad hurtownią danych dla systemu IACS zostały rozpoczęte w lipcu 2001 roku. W czasie realizacji produktu wielokrotnie zmieniały się założenia funkcjonalne, architektura systemu oraz organizacja prac projektowych.

Obecnie hurtownia IACS gromadzi dane pochodzące z rejestru gospodarstw rolnych, rejestru zwierząt oraz rejestru ZSZiK (Zintegrowany System Zarządzania i Kontroli). Kolejny etap prac to przystosowanie hurtowni do obsługi schematów obszarowych. Hurtownia IACS została zaprojektowana i zrealizowana przez firmę Pentacomp. Rozwiązanie zostało zaimplementowane na serwerze bazy danych Oracle 9i z wykorzystaniem oprogramowania typu ETL oraz narzędzi OLAP firmy Oracle. Dane w hurtowni gromadzone są w cyklu dobowym, a raporty udostępniane w trybie online.

W referacie przedstawiona zostanie historia projektu, główne problemy organizacyjne oraz technologiczne, na jakie napotkano przy jego realizacji oraz zastosowane rozwiązania.

## Informacja o autorze:

Absolwent Instytutu Informatyki Politechniki Białostockiej. Od sierpnia 1997 roku pracuje dla firmy Pentacomp Systemy Informatyczne. W latach 1997-1999 jako projektant-programista brał udział w realizacji systemu IKSS (Infrastruktura Krytyczna Systemu Śledzenia wagonów i przesyłek) realizowanego na zlecenie PKP, w latach 1999-2001 był analitykiem-projektantem korporacyjnej hurtowni danych CIS (Company Information System) w PTC, od 2001 kierownik oraz główny projektant części analitycznej systemu IACS dla Agencji Restrukturyzacji i Modernizacji Rolnictwa (ARiMR).

## 1. Wprowadzenie

W realizacji dużych projektów informatycznych dla instytucji publicznych ma miejsce powtarzanie się pewnych charakterystycznych zjawisk. Często są to przedsięwzięcia zakrojone na szeroką skalę, ich realizacja przebiega powolnie, odbierane systemy informatyczne nie spełniają oczekiwań klienta, co w konsekwencji prowadzi do tego, że część tych projektów kończy się niepowodzeniem.

W artykule przedstawione zostały potencjalne cele budowy systemów analitycznych w sektorze administracji publicznej, stawiane wymagania w zakresie wykorzystania zgromadzonych danych oraz korzyści jakie przynosi zastosowanie rozwiązań opartych na hurtowni danych.

W drugiej części artykułu, korzystając z doświadczeń nabytych podczas realizacji hurtowni danych systemu IACS dla Agencji Restrukturyzacji i Modernizacji Rolnictwa, podjęto próbę zidentyfikowania charakterystycznych problemów, jakie mogą wystąpić przy realizacji tego typu przedsięwzięć oraz określenia sposobów ich rozwiązywania.

W rozdziale 4 bazując na dotychczasowej praktyce z wykorzystaniem rozwiązań Oracle, dokonano oceny technologii i narzędzi, istotnych z punktu widzenia realizacji hurtowni danych.

## 2. Założenia realizacji hurtowni danych w administracji publicznej

### Wymagania stawiane realizowanym systemom

W obecnych czasach firmy komercyjne dążą do zwiększenia opłacalności prowadzonej działalności, pozyskania większej liczby klientów oraz zapewnienia sobie przewagi nad konkurencją. Do realizacji tych celów niezbędne jest zbieranie, sprawny przepływ oraz analizowanie gromadzonych informacji, w czym pomagają odpowiednie rozwiązania informatyczne.

Inne podłoże mają systemy informatyczne tworzone w sektorze administracji publicznej. Podstawowe zadania stojące przed instytucjami państwowymi to wykorzystanie technologii informatycznych do:

- usprawnienia realizacji powierzonych im funkcji społecznych,
- zapewnienia odpowiedniego poziomu usług,
- obniżenia kosztów działania instytucji.

Podstawowe cele funkcjonalne stawiane realizowanym systemom informatycznym dotyczą zapewnienia optymalnego przepływu informacji z zakresu działania administracji publicznej. Należą do nich:

- rejestracja, kontrola i przetwarzanie danych związanych z prowadzonymi działaniami w ramach realizacji funkcji publicznych,
- utrzymywanie i dostęp do danych referencyjnych,
- zapewnienie wymiany informacji pomiędzy poszczególnymi jednostkami instytucji,
- dostarczenie informacji do celów zarządzania,
- udostępnienie informacji dla innych instytucji oraz społeczeństwa.

Wymagania niefunkcjonalne stawiane realizowanym systemom to przede wszystkim:

- zagwarantowanie dużej dostępności, jakości i bezpieczeństwa danych, w tym spełnienie wymogów ustawy o ochronie danych osobowych,
- zagwarantowanie odpowiedniej wydajności i niezawodności systemu gromadzącego ogromne ilości danych, w którym czas wykonania poszczególnych działań jest szczególnie istotny,
- elastyczność i skalowalność systemu informatycznego umożliwiającą szybkie dostosowywanie się do zmian prawnych, w tym dostosowywanie się do prawa Unii Europejskiej, zmian organizacyjnych oraz technologicznych,
- otwartość umożliwiającą wymianę informacji pomiędzy różnymi organizacjami.

Systemy, które mają sprostać wymienionym wymaganiom, są często dedykowanymi rozwiązaniami realizowanymi na konkretne potrzeby wynikające z profilu prowadzonej działalności instytucji. Podstawą ich budowy są odpowiednie przepisy i rozporządzenia prawne.

Często są to zintegrowane systemy klasy OLTP (*ang. On-Line Transaction Processing*) wspierające bieżącą działalność operacyjną organizacji. Systemy te są nastawione na obsługę wielu krótkich i prostych transakcji (tj. rejestracja dokumentu, obsługa klienta), projektowane są pod kątem uzyskania maksymalnej wydajności dotyczącej odczytu i modyfikacji pojedynczych zapisów informacji przechowywanych w bazie danych systemu. Taki typ systemu nie wspomaga jednak procesów analizy danych, w których najistotniejsze jest uzyskiwanie zbiorczych informacji oraz wykonywanie złożonych zapytań na dużej liczbie przechowywanych danych. Są to cechy charakterystyczne dla systemów wspomagania decyzji (*ang. Decision Support Systems - DSS*).

## Cele budowy systemów analitycznych

Do realizacji potrzeb analitycznych administracji tj. dostarczenie informacji do celów zarządzania i kontroli, udostępnienie zbiorczych zestawień statystycznych innym organom administracji publicznej tworzy się dedykowane rozwiązania klasy DSS.

Podstawowym elementem systemu jest składnica danych (*ang. data warehouse*), będąca rejestrem gromadzącym spójne, zintegrowane i historyczne dane pochodzące z rejestrów systemów OLTP, jak również dodatkowe dane zewnętrzne (np. arkusze kalkulacyjne, pliki tekstowe) niezbędne dla realizacji potrzeb analitycznych organizacji.

Realizowanym systemom analitycznym często stawia się następujące wymagania:

- zapewnienie krótkiego czasu wykonania analiz pozwalającego na pracę interakcyjną
- dostarczenie środowiska analitycznego, które jest parametryzowane, otwarte na nowe potrzeby intuicyjne w obsłudze, udostępniające złożone formy prezentacji i analizy danych
- utrzymanie danych szczegółowych wraz z historią zmian wprowadzanych w danych.

## Architektura systemu hurtowni danych

Na rysunku 1 została przedstawiona typowa ogólna architektura systemu wspomagania decyzji. Podstawowym elementem prezentowanego systemu jest składnica danych gromadząca i integrująca dane pochodzące z różnych źródeł danych. W przedstawionym rozwiązaniu hurtowni danych zostały wyróżnione dwa typu rejestrów:

- **składnica danych detalicznych** – podstawowy rejestr gromadzący dane ze wszystkich niezbędnych źródeł danych na poziomie szczegółowym z utrzymaniem historii zmian we wprowadzanych danych; struktura rejestru jest projektowana pod kątem zapewnienia dużej wydajności w dostępie do danych,

- **składnica danych tematycznych** – rejestry danych zorientowane tematycznie na poszczególne obszary działalności organizacji (*ang. data marts*), często projektowane tak, aby przechowywały dane zagregowane.

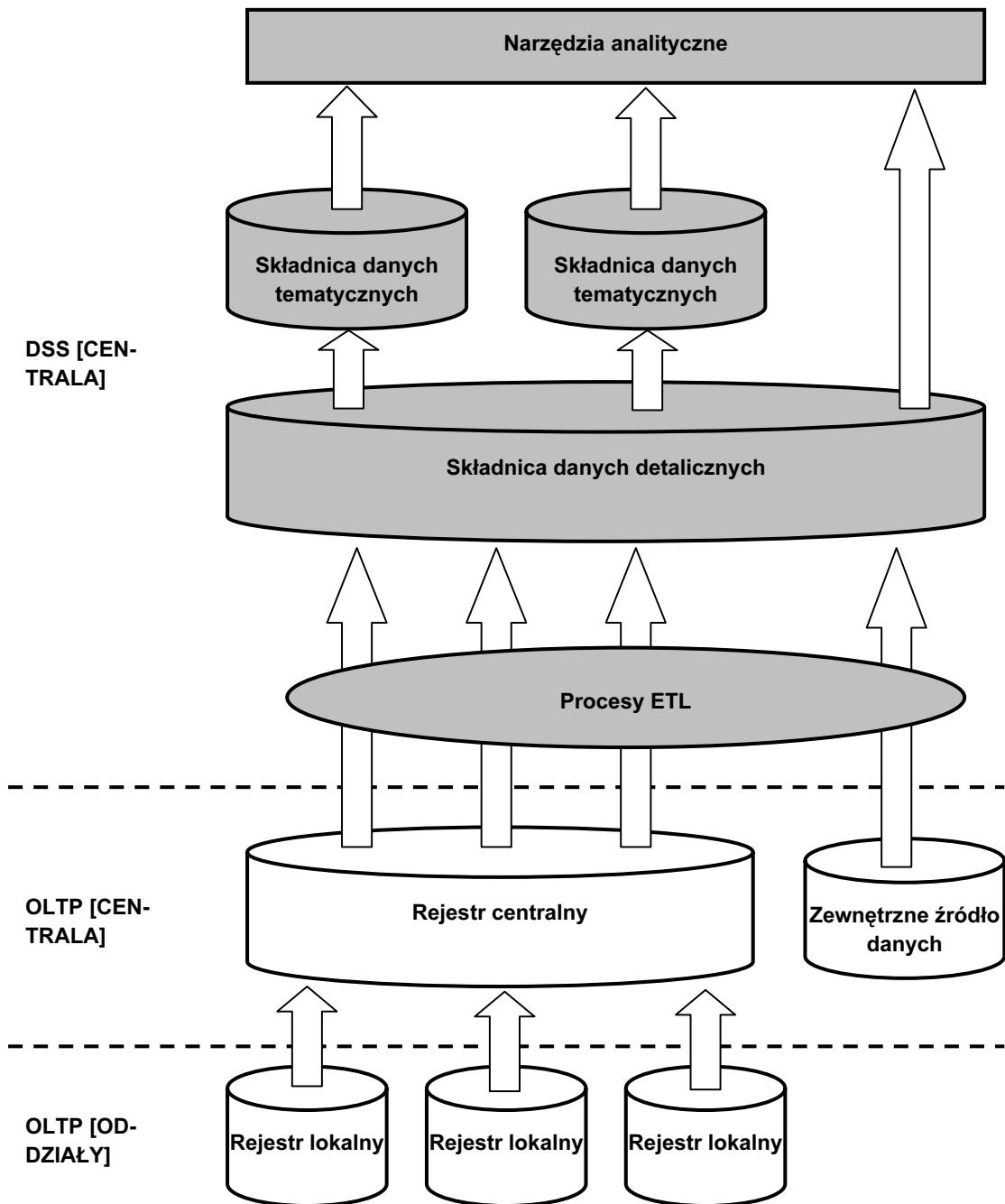
Często, z różnych przyczyn – technologicznych, organizacyjnych czy finansowych, rozwiązania tworzone w rzeczywistości odbiegają od prezentowanego schematu. Niekiedy realizowane są hurtownie na konkretne potrzeby w postaci rejestrów danych tematycznych, bez wspólnego obszaru danych detalicznych. Innym, często spotykanym rozwiązaniem jest zbudowanie jedynie składnicy danych detalicznych, w najprostszej wersji może być to kopia bazy części operacyjnej do celów raportowania (*ang. operational data store - ODS*).

Kolejnym istotnym elementem systemu są **procesy ETL** (*ang. Extraction, Transformation, Loading*), odpowiedzialne za cykliczne zasilanie hurtowni nowymi danymi pochodzącymi z systemów źródłowych. Głównym zadaniem tych procesów jest:

- ekstrakcja danych z rejestrów źródłowych (systemów OLTP oraz dodatkowych źródeł danych),
- przekształcenie danych do postaci struktur docelowych,
- wyczyszczenie danych wprowadzanych do hurtowni, w tym eliminacja błędów niespójności, niekompletności, redundancji, wraz z rejestracją wykrytych nieprawidłowości w danych,
- załadowanie poprawnych danych do hurtowni.

Wsparciem dla ETL jest odpowiednie środowisko do cyklicznego uruchamiania, zarządzania i monitorowania stanu procesów. Podstawą jego działania są metadane opisujące poszczególne procedury, relacje pomiędzy nimi, reguły transformacji i czyszczenia danych oraz parametry sterujące cyklicznym uruchamianiem zadań.

Ostatnim elementem przedstawionego rozwiązania są **narzędzia analityczne** umożliwiające prowadzenie analiz danych przechowywanych w hurtowni. W zależności od potrzeb analitycznych stosuje się różne typy narzędzi do analizy danych: od prostych programów raportujących, poprzez narzędzia OLAP (*On-Line Analytical Processing*) umożliwiające interaktywną analizę danych i różne formy prezentacji wyników, do złożonych systemów wyszukujących wzorców i regularności w dużych zbiorach danych z zastosowaniem zaawansowanych metod statystycznych lub sztucznej inteligencji (*ang. data mining*).



Rys. 1. Przykładowa architektura systemu wspomaganie decyzji

### Zalety realizacji systemu hurtowni danych

System wspomaganie decyzji nastawiony jest na spełnienie oczekiwań analitycznych, wynikających zarówno z wewnętrznych potrzeb organizacji jak i narzuconych lub wymaganych przez zewnętrzne organy administracji.

Zalety przedstawionej architektury systemu z wykorzystaniem hurtowni danych to przede wszystkim:

- krótki czas wykonywania zapytań i analiz, który umożliwia pracę interaktywną, osiągnięty poprzez odpowiednie zdefiniowanie struktur danych i przygotowane niezbędnych danych zbiorczych,
- duża wydajność w dostępie do danych osiągnięta poprzez oddzielenie procesów analitycznych od operacyjnego działania systemu,
- wysoka jakość odczytywanych informacji, będąca wynikiem transformacji, weryfikacji poprawności i czyszczenia danych pochodzących z systemów źródłowych,
- łatwo dostępna baza informacji historycznych oraz danych zbiorczych do celów analitycznych i statystycznych,
- otwartość systemu na integrację nowych źródeł danych oraz podstawa do rozbudowy systemów analitycznych i statystycznych tj. OLAP, data mining.

### **3. Problemy realizacji hurtowni danych w administracji publicznej**

Hurtownia danych obejmuje całą działalność instytucji. Jest to złożone i kosztowne rozwiązanie, w którym podczas realizacji występuje wiele zagrożeń, które mogą w konsekwencji doprowadzić do niepowodzenia przedsięwzięcia.

Poniżej zostały opisane typowe problemy jakie mogą wystąpić podczas realizacji systemów hurtowni danych w sektorze administracji publicznej oraz wskazówki, które mogą wpłynąć na zmniejszenie ryzyka niepowodzenia tego typu przedsięwzięcia.

#### **Niespełnienie oczekiwań końcowych użytkowników systemu**

Podczas realizacji projektu bardzo istotna jest świadomość odbiorców systemu co do informacji, formy dostępu i prezentacji danych pozyskiwanych z hurtowni. Odbiorcy powinni być świadomi korzyści wynikających z uzyskiwania tych informacji. Zapobiega to realizacji systemu, z którego można odczytać jedynie nikomu niepotrzebne rzeczy.

Głównym adresatem systemu hurtowni jest kadra zarządzająca wyższego szczebla, odpowiedzialna za podejmowanie decyzji w ramach organizacji, mająca największą świadomość strategii działania i potrzeb z tego wynikających. To przede wszystkim te osoby powinny być zaangażowane w aktywną współpracę z wykonawcą na etapie doprecyzowywania wymagań dla systemu. W praktyce jednak kadra zarządcza, zbyt zajęta innymi sprawami, deleguje do współpracy z dostawcą rozwiązania kadrę niższego szczebla lub komórkę IT, które nie mają wielkiej świadomości w zakresie oczekiwanych wyników.

Dobłą praktyką jest uzgadnianie z odbiorcą systemu szczegółowego zakresu informacyjnego i specyfikacji tworzonego systemu na etapie analizy wymagań. Na tym etapie pomocna jest realizacja prototypów funkcjonalnych, umożliwiającą lepszą weryfikację potrzeb klienta.

Podczas realizacji projektu hurtowni danych należy stosować podejście iteracyjne w zakresie projektowania, implementacji i wdrażania systemu. Przy odbiorze kolejnych przyrostów funkcjonalności hurtowni, użytkownicy pogłębiają swoją wiedzę na temat systemów analitycznych, chętniej definiują swoje potrzeby, widząc jakie korzyści daje wprowadzane rozwiązanie.

Spełnienie oczekiwań to także dobór odpowiednich narzędzi analitycznych. W zależności od wymagań końcowego odbiorcy mogą być to proste narzędzia dostarczające gotowe raporty, narzę-

dzia umożliwiające interaktywną analizę danych, bądź też bardziej skomplikowane rozwiązania umożliwiające pozyskiwanie informacji z danych zgromadzonych w hurtowni.

### **Masowy dostęp do danych hurtowni**

W systemach przeznaczonych dla administracji publicznej często mamy do czynienia z danymi, które powinny być dostępne dla szerokiego kręgu odbiorców. Jeżeli wynika to z mocy prawa, możliwe jest nawet, że dane powinny być dostępne publicznie, choć z reguły sprowadza się to do udostępniania informacji różnorodnym instytucjom. Powoduje to, że w odróżnieniu od hurtowni korporacyjnych, mamy tu do czynienia z masowym dostępem do danych. Zapewnienie takiego dostępu jest niekiedy poważnym problemem projektowym, ponieważ wymaga to rozwiązania wielu dodatkowych problemów związanych z wydajnością, niezawodnością systemu, bezpieczeństwem danych, czy dodatkowymi przepisami prawa (np. ustawą o ochronie danych osobowych).

### **Zmienność założeń funkcjonalnych podczas realizacji systemu**

W czasie realizacji złożonych i długotrwałych projektów często dochodzi do zmian założeń dla systemu. W sektorze administracji publicznej źródłem zmian są zazwyczaj decyzje „polityczne” bądź też aktualizacje przepisów prawnych.

Zmiany założeń pociągają za sobą zmiany w realizowanym projekcie, mogą dotyczyć systemów będących źródłem danych dla hurtowni jak i też środowiska analitycznego. Zdarza się, że dotychczas zbudowane fragmenty systemu należy wyrzucić do kosza i rozpocząć ich realizację od nowa.

Zazwyczaj zmian założeń realizacyjnych nie da się uniknąć, można natomiast próbować minimalizować koszty ich wprowadzenia. Sposobem ograniczenia ryzyka jest zastosowanie wieloetapowej sekwencyjnej realizacji systemu, w której doprecyzowanie następnych wymagań następuje po zakończeniu realizacji kolejnego etapu.

Na początku prac konieczne jest zdefiniowanie szczegółowej architektury tworzonej hurtowni oraz opracowanie standardów projektowych dotyczących realizacji poszczególnych komponentów. Opracowanie i udokumentowanie ogólnych założeń realizacyjnych zapewnia utrzymanie spójności tworzonego w kolejnych iteracjach rozwiązania oraz zmniejsza koszty działań związanych z wprowadzaniem nowej funkcjonalności.

Kolejnym, bardzo istotnym czynnikiem ograniczającym ryzyko jest zastosowanie odpowiednich narzędzi spełniających funkcję repozytorium prowadzonych prac projektowych w zakresie wsparcia w definiowaniu i zarządzaniu zmianami struktur projektowych, jakie definiuje się w kolejnych etapach realizacji systemu.

### **Niedostępność źródeł danych**

Tworząc hurtownię danych dla instytucji mamy do czynienia z dużą liczbą zewnętrznych źródeł danych. Charakteryzują się one różną strukturą, często są to systemy słabo udokumentowane, a niekiedy znajdują się one dopiero na etapie projektowania.

Podczas realizacji projektu należy duży nacisk położyć na przeprowadzenie szczegółowej analizy danych źródłowych. Na tym etapie należy dokładnie zapoznać się z istniejącymi specyfikacjami systemów źródłowych, przeprowadzić dodatkowe rozmowy z autorami lub osobami, które najlepiej znają te systemy, przejrzeć i przeanalizować dane jakie są w nich rejestrowane. Wynikiem przeprowadzonej analizy danych powinien być dokument zawierający opis i charakterystykę pojedynczych atrybutów składowanych w poszczególnych systemach źródłowych istotnych z punktu widzenia projektowanej hurtowni danych. Takie opracowanie stanowi podstawę do stworzenia docelowego modelu hurtowni oraz zaprojektowania prawidłowych procedur zasilających

(ETL). Jest to także podstawa prac podczas rozbudowy hurtowni o nowe źródła danych oraz wprowadzania zmian wynikających z aktualizacji systemów źródłowych.

W praktyce występują przypadki, kiedy system hurtowni danych realizowany jest pod nieistniejące bądź też równolegle tworzone systemy źródłowe. Taki stan może być na przykład konsekwencją odpowiednich przepisów prawa. Jest to sytuacja bardzo nietypowa dla projektu hurtowni danych. Tworząc ją równolegle do systemów źródłowych mamy okazję wpłynięcia na decyzje projektowe realizowanego systemu źródłowego. Wtedy poprzez zagwarantowanie wykorzystania odpowiednich mechanizmów w bazie źródłowej np. odpowiednich formatów danych, kolumn audytu w tabelach, indeksów czy też dzienników zmian, wpływamy na zwiększenie efektywności i jakości zasilania hurtowni danymi.

W przypadku, gdy hurtownia danych jest tworzona pod nieistniejący jeszcze system źródłowy, może się zdarzyć, że opracowanie interfejsu wymiany danych pomiędzy systemami jest elementem projektu hurtowni, a budowany w przyszłości system będzie się musiał do niego dostosować.

## Niska jakość danych wejściowych

Mnogość i różnorodność systemów źródłowych to częsta cecha występująca w administracji. Realizując projekt możemy mieć do czynienia z wieloma różnego typu rejestrami terenowymi, które powstawały w dalekiej przeszłości, kiedy nie myślano o tworzeniu dużych i jednolitych rozwiązań informatycznych. Często okazuje się, że w rejestrach źródłowych wiele informacji jest powielanych, dane zawierają błędy, zaimplementowane w systemach reguły walidacji dopuszczają wprowadzanie danych w różnych formatach.

Chcąc zapewnić w hurtowni odpowiednią jakość uzyskiwanych informacji należy poświęcić dużo pracy na określenie odpowiednich reguł transformacji i czyszczenia danych źródłowych oraz zdefiniowanie procesów, odpowiedzialnych za ich przetwarzanie. Istotny jest w tym dobór odpowiednich narzędzi, które wspierają projektowanie procedur i procesów ETL, jak również zapewniają wsparcie podczas ich cyklicznego uruchamiania w zakresie monitorowania i kontroli procesu, obsługi sytuacji błędnych czy też raportowania wyników zasilania hurtowni.

## Problemy wydajnościowe

Często dopiero na etapie wdrożenia rozwiązania hurtowni danych zaczynają ujawniać się problemy związane z niemożliwością załadowania danych źródłowych lub zbyt długimi czasami odpowiedzi na wysłane do bazy zapytania. Problemy tego rodzaju nie są oczywiście charakterystyczne wyłącznie dla hurtowni danych dla administracji, ale tutaj mogą się ujawniać w dużo większym stopniu. Wynika to w szczególności z dużego lub bardzo dużego wolumenu danych, jeśli system ma obejmować cały kraj (a więc np. w „najgorszym” wypadku 38 milionów obywateli), a ponadto z bardzo różnej charakterystyki danych (np. zupełnie inaczej zachowują się dane dla Warszawy, czy aglomeracji śląskiej, a inaczej dla województwa Podkarpackiego, czy Podlaskiego). Może to powodować poważne trudności w dobraniu odpowiedniej charakterystyki danych testowych, szczególnie w sytuacji, gdy rzeczywiste dane źródłowe nie są dostępne.

Dlatego też prace nad zapewnieniem odpowiedniej wydajności rozwiązania należy rozpocząć już w fazie projektowania ogólnej architektury systemu. Na tym etapie należy zebrać wymagania pojemnościowe dla tworzonego systemu, opracować prawidłowy model bazy danych oraz procesy zasilania hurtowni. Należy również zastanowić się nad wykorzystaniem odpowiednich mechanizmów systemu zarządzania bazą danych oraz doбором właściwych narzędzi, które umożliwią efektywne zasilanie oraz odczyt danych.

Jednak nie wszystkie problemy związane ze zbyt długimi czasami wykonania operacji można rozwiązać na etapie projektowania systemu. Dlatego też ważnym etapem realizacji jest przeprowadzanie testów wydajnościowych. Uruchamianie i monitorowanie pracy systemu operującego na

dużych zbiorach danych pozwala zidentyfikować miejsca występowania problemów i jest podstawą do jego strojenia. W tym przypadku kluczową rolę odgrywają wykorzystywane narzędzia i ich możliwości w zakresie optymalnego wykorzystania zasobów systemu operacyjnego oraz bazy danych.

## 4. Wykorzystanie technologii Oracle do budowy hurtowni danych

Opisywane w poprzednim rozdziale zagrożenia występujące podczas realizacji złożonych systemów hurtowni danych często można minimalizować decydując się na dobór technologii i narzędzi informatycznych.

W niniejszym rozdziale, bazując na dotychczasowych doświadczeniach w tworzeniu dużych systemów hurtowni danych, a szczególnie systemu IACS, scharakteryzowane zostały produkty firmy Oracle stosowane do realizacji systemów DSS, pod kątem ich użyteczności do tego celu, ze szczególnym nastawieniem na:

- zapewnienie odpowiedniej wydajności tworzonego rozwiązania,
- zapewnienie odpowiedniej funkcjonalności w zakresie obsługi procesów ETL,
- spełnienie oczekiwań docelowych użytkowników systemu analitycznego,
- wsparcie dla prowadzonych prac projektowych w zakresie modelowania rozwiązania i zarządzania zmianami w projekcie.

### Baza danych Oracle 9i

Wykorzystanie systemu zarządzania bazą danych Oracle 9i daje duże możliwości w zakresie opracowania efektywnego ładowania, składowania oraz sprawnego dostępu do dużych wolumenów danych. Poniżej zostały opisane mechanizmy, których zastosowanie wpływa na zwiększenie wydajności tworzonego rozwiązania.

#### Partycjonowanie

Wykorzystanie mechanizmu partycjonowania umożliwia dzielenie na mniejsze części dużych struktur przechowywanych w bazie danych. Podczas wykonywania poleceń optymalizator bazy danych analizuje zapytanie w celu wyeliminowania z procesu wyszukiwania tych partycji, które nie zawierają wyszukiwanych informacji, następnie wykonanie operacji odbywa się tylko na wybranych podzbiorach danych. W ten sposób skraca się czas wykonania zapytania, a co za tym idzie zwiększa się wydajność rozwiązania.

W przypadku wykorzystywania mechanizmu należy zwracać szczególną uwagę na konstruowanie odpowiednich zapytań odwołujących się do partycjonowanych tabel. Często tabele dzieli się na partycje wg kilku zależnych od siebie kolumn. Przykładem może być podział administracyjny kraju składający się z następujących zależnych od siebie i unikalnych atrybutów: kod województwa, kod powiatu, kod gminy, kod miejscowości. Zbudowanie zapytania z pominięciem w warunku selekcji elementu nadrzędnego np. kodu województwa powoduje, że zamiast pojedynczej partycji przeszukiwana jest cała tabela.

#### Indeksowanie

W zapewnieniu szybkiego dostępu do danych pomagają indeksy zakładane na tabelach bazy danych. Baza Oracle umożliwia tworzenie standardowych indeksów o strukturze B-drzewa, indeksów bitmapowych, tabel o strukturze indeksu oraz indeksów w oparciu o funkcje działające na jednej lub większej liczbie kolumn indeksowanej tabeli.

Decyzję o wyborze właściwego typu indeksu należy podejmować na podstawie oceny liczności, kardynalności oraz sposobu wykorzystania poszczególnych kolumn bazy danych.

Przykładowo, indeksy bitmapowe stosuje się zazwyczaj dla kolumn, które cechują się małą kardynalnością. Kardynalność jest jednak uzależniona od liczności całego zbioru. Doświadczenie pokazuje, że nawet w przypadku, jeżeli istnieje około 100.000 różnych wartości, ale cały zbiór liczy ponad 1.000.000 rekordów, zastosowanie indeksu bitmapowego jest bardziej wydajne od standardowego indeksu o strukturze B-drzewa.

Duże zastosowanie w projektowaniu hurtowni mają również indeksy działające w oparciu o funkcje na kolumnach typu DATE. Jeżeli z bazy próbujemy uzyskać dane dotyczące danego dnia, często w warunku zapytania wykorzystujemy funkcję `TO_CHAR(date, format)` lub też `TRUNC(date)`.

### Optymalizacja zapytań

Tworząc system można spróbować zaufać wbudowanym mechanizmom optymalizacji wykonywania zapytań. Jednakże często takie podejście okazuje się niewystarczające. Złożone zapytania operujące na dużych zbiorach danych zazwyczaj wymagają procesu strojenia. Tu z pomocą przychodzi rozwiązanie bazy danych Oracle 9i, które umożliwiają wybór trybu pracy optymalizatora, pozwalają na zbieranie statystyk opisujących poszczególne zbiory danych oraz prowadzenie analizy planu wykonania poszczególnych operacji, a także poprzez użycie wskazówek (*ang. hints*) dają możliwość zdefiniowania własnego planu wykonania dla zapytania.

### Perspektywy zmaterializowane

Perspektywy zmaterializowane (*ang. materialized views*) umożliwiają składowanie wyników zapytań w postaci tabel bazy danych.

Jedną z podstawowych technik stosowanych w hurtowniach danych dla zwiększenia wydajności jest tworzenie agregatów. Tworzone agregaty zawierają zazwyczaj sumaryczne dane pochodzące z jednej lub wielu połączonych tabel przechowujących dane szczegółowe. W bazie danych Oracle można do tego celu wykorzystać perspektywy zmaterializowane oraz mechanizm przepisania zapytań (*ang. query rewrite*). Po zastosowaniu tych mechanizmów, końcowy użytkownik wykonując zapytania odwołując się do tabel detalicznych może uzyskać szybką odpowiedź pochodzącą ze zdefiniowanej perspektywy. Dzieje się tak na skutek działania optymalizatora, który analizuje składnię wysłanego zapytania SQL i w przypadku zgodności z definicją perspektywy zmaterializowanej, nie odwołuje się do tabel detalicznych, lecz zwraca wyniki przechowywane w perspektywie.

Oczywisty problem, który pojawia się w sytuacji wykorzystywania perspektyw zmaterializowanych dotyczy aktualności przechowywanych tam danych. Dostępna jest rozbudowana funkcjonalność umożliwiająca odświeżanie danych zawartych w perspektywach na podstawie zmian zachodzących w tabelach będących źródłem danych perspektywy. Perspektywa może być odświeżana asynchronicznie – cyklicznie lub na żądanie użytkownika albo synchronicznie – wtedy, gdy pojawią się zmiany danych źródłowych. Podczas odświeżania można zdecydować się na przeładowanie wszystkich rekordów perspektywy lub też wybrać metodę szybkiego odświeżania, która bazując na zmianach tabel źródłowych, ładuje tylko te wiersze, które zostały zmienione od czasu ostatniej aktualizacji. Z opisanych cech wynika, że perspektywy zmaterializowane mogą również znaleźć zastosowanie podczas procesów inkrementalnego zasilania hurtowni danych.

Decydując się na wykorzystanie mechanizmu perspektyw zmaterializowanych oraz przepisania zapytań należy jednak wykazać się pewną ostrożnością. Po zapoznaniu się z ich szczegółowymi cechami i ograniczeniami, może się na przykład okazać, że niemożliwe jest przepisanie

zapytań zawierających klauzulę CONNECT BY, bądź też nie są wykorzystywane wskazówki dla optymalizatora podczas odwoływania się w zapytaniu do bazy zdalnej.

### Mechanizmy wspierające procesy aktualizacji dużych porcji danych

System Oracle 9i udostępnia pewne rozwiązania, na które warto zwrócić uwagę przy realizacji procesów ETL. Są to:

- polecenie MERGE – funkcja realizująca wstawienie lub aktualizację wierszy wywoływana za pomocą jednego polecenia, zastępuje wywołanie dwóch poleceń: INSERT i UPDATE, zapobiega w ten sposób niepotrzebnemu przełączaniu kontekstu pomiędzy PL/SQL a SQL, co skutkuje wzrostem wydajności przetwarzania.
- wielotabelowe operacje INSERT – pojedynczym poleceniem SQL można jednocześnie wstawić nowe danych do kilku tabel, co wpływa na zwiększenie wydajności, gdyż nie wymaga wielokrotnego przeszukiwania danych źródłowych.
- mechanizm BULK BINDS – w przypadku sekwencyjnego przetwarzania dużych zbiorów danych mechanizm minimalizuje liczbę przełączeń kontekstu pomiędzy PL/SQL i SQL poprzez odłożenie w kolekcji wykonanych operacji, a następnie hurtowe zapisanie fizycznych zmian do bazy danych.

Wykorzystanie opisanych mechanizmów może znacząco skrócić czas przetwarzania danych. Przeprowadzone testy pokazują na przykład, że przy wykorzystaniu mechanizmu MERGE można zaoszczędzić nawet do 50% czasu w porównaniu z procedurą opartą o polecenia INSERT i UPDATE. Podczas wyboru narzędzia wspierającego projektowanie procesów ETL warto zwrócić uwagę, czy umożliwi ono generację kodu z wykorzystaniem opisanych mechanizmów.

### Rozwiązania wspierające prowadzenie analiz danych

Podczas projektowania hurtowni należy wybrać odpowiedni model przetwarzania. Można się zdecydować na przetwarzanie struktur relacyjnych bazy danych (*ang. Relational On-Line Analytical Processing – ROLAP*) zamodelowanych w postaci schematu gwiazdy (*ang. star*) lub płatką śniegu (*ang. snowflake*).

Alternatywą tego rozwiązania jest zastosowanie analitycznego przetwarzania struktur wielowymiarowych (*ang. Multidimensional On-Line Analytical Processing – MOLAP*). Wsparciem bazy danych w tym zakresie jest Oracle OLAP (*ang. On-Line Analytical Processing*). Umożliwia on definiowanie wymiarów i faktów przechowywanych w bazie danych, jest silnikiem wspomagającym obliczenia wielowymiarowych danych, umożliwia tworzenie kostek wielowymiarowych oraz zawiera funkcje wspierające wykonywanie operacji analitycznych.

W praktyce, kiedy mamy do czynienia z bardzo dużymi zbiorami danych, które chcemy analizować w ramach wielu różnych wymiarów, budowanie kostek wielowymiarowych może być niemożliwe z powodu ograniczeń pojemnościowych i wydajnościowych systemu. W takim przypadku zazwyczaj podejmuje się decyzje realizacji relacyjnej hurtowni danych.

Alternatywnym rozwiązaniem jest układ hybrydowy (*ang. Hybrid On-Line Analytical Processing – HOLAP*), w którym część elementów zostaje zaprojektowana w modelu relacyjnym, a część w modelu wielowymiarowym. Przy wyborze rozwiązania typu HOLAP należy dobrać odpowiednie narzędzia analityczne, które będą potrafiły przetwarzać zarówno dane relacyjne jak i wielowymiarowe.

## Oracle Warehouse Builder

Oracle Warehouse Builder jest narzędziem do projektowania hurtowni danych. Narzędzie umożliwia definiowanie źródeł danych, projektowanie struktur docelowych hurtowni oraz tworzenie procedur transformacji danych.

Narzędzie umożliwia graficzne modelowanie procedur transformacji danych. Posiada bogaty zestaw operatorów mapowania, dzięki którym można zdefiniować zazwyczaj większość wymaganych przekształceń. W praktyce jednak zdarzają się również przypadki, gdy mamy do czynienia z bardzo złożonymi regułami transformacji danych i wtedy często okazuje się, że nawet najlepsze rozwiązania graficzne nie zastąpią bezpośredniej implementacji procedury.

Tworzone za pomocą Oracle Warehouse Builder procedury transformacji są optymalizowane pod kątem wydajnościowym. Podczas projektowania we właściwościach mapowania można między innymi określić tryb, w jakim ma być wykonywane przetwarzanie (operujące na całym zbiorze lub pojedynczych rekordach) oraz wskazówki dla optymalizatora bazy danych. Generowany na podstawie projektu kod wykorzystuje rozwiązania bazy danych zwiększające wydajność przetwarzania, tj. mechanizm BULK BINDS, instrukcję MERGE oraz wielotabelową operację INSERT.

Oracle Warehouse Builder dostarcza mechanizm pozwalający na śledzenie wykonywania operacji. Dzięki temu możliwe jest szybkie wykrycie nieprawidłowości, jakie pojawiły się w danych źródłowych podczas procesu zasilania hurtowni danych. Mechanizm śledzenia automatycznie tworzy raporty zawierające szczegółowe informacje dotyczące wykrytych błędów.

Oracle Warehouse Builder umożliwia zarządzanie zmianami w strukturach wejściowych i docelowych. Różnice występujące pomiędzy aktualnym projektem, a definicją zewnętrznych tabel są wykrywane przez narzędzie i możliwe do automatycznego wprowadzenia do repozytorium. Często jednak zmianie ulegają typy kolumn, ich znaczenie biznesowe lub też pewne atrybuty są usuwane ze źródeł, co może znacząco wpłynąć na logikę przetwarzania. W takich przypadkach nie pozostaje zazwyczaj nic innego, jak przejrzanie i manualne wprowadzenie zmian we wszystkich zaprojektowanych mapowaniach.

## Oracle Discoverer

Oracle Discoverer to rozwiązanie umożliwiające definiowanie i prowadzenie analiz przekrojowych na podstawie zdefiniowanych struktur. Na produkt w architekturze klient-serwer składają się dwie aplikacje: Oracle Discoverer Administrator oraz Oracle Discoverer Desktop. W trójwarstwowej architekturze internetowym odpowiednikiem narzędzia Oracle Discoverer Desktop jest Oracle Discoverer Plus.

### Oracle Discoverer Administrator

Oracle Discoverer to rozwiązanie umożliwiające definiowanie i prowadzenie analiz przekrojowych na podstawie zdefiniowanych struktur. Narzędzie to może pracować w dwóch architekturach: klient-serwer i trójwarstwowej. Na produkt w architekturze klient-serwer składają się dwie aplikacje: Oracle Discoverer Administrator oraz Oracle Discoverer Desktop. W trójwarstwowej architekturze internetowym odpowiednikiem narzędzia Oracle Discoverer Desktop jest Oracle Discoverer Plus.

Dużą zaletą narzędzia jest możliwość optymalizowania czasu dostępu do bazy danych poprzez:

- tworzenia podsumowań w postaci perspektyw zmaterializowanych, które mogą być następnie wykorzystane podczas przepisywania zapytań,
- definiowanie folderów opartych na zapytaniach SQL oraz ustaleniu wskazówek dla optymalizatora bazy danych,

- wykorzystanie w definicjach folderów funkcji analitycznych bazy danych.

Oracle Discoverer jest rozwiązaniem typu ROLAP, a co za tym idzie nie daje możliwości wykorzystania faktów, wymiarów oraz wielowymiarowych struktur bazy danych zdefiniowanych przy pomocy Oracle OLAP.

### **Oracle Discoverer Desktop (Oracle Discoverer Plus)**

Oracle Discoverer Desktop jest narzędziem przeznaczonym dla końcowego użytkownika, służącym do prowadzenia analiz danych. Na podstawie struktur zdefiniowanych przy pomocy narzędzia Oracle Discoverer Administrator użytkownik może wybrać interesujące go dane, przeprowadzić ich analizę oraz przedstawić wyniki w formie tabelarycznej lub graficznej.

W ramach prowadzenia interaktywnej analizy danych, użytkownik może korzystać z typowej dla rozwiązań typu OLAP funkcjonalności, tj. drążenie w górę, w dół oraz do powiązanych elementów, tworzenie ograniczeń, podsumowań, dodatkowych obliczeń oraz definiowanie sposobu sortowania danych.

Jest to stosunkowo proste w obsłudze i rozbudowane funkcjonalne narzędzie, jednak przy realizacji bardzo skomplikowanych obliczeń wymagające od użytkownika dobrej znajomości funkcji analitycznych udostępnianych przez Oracle, a także składni poleceń SQL.

## **5. Podsumowanie**

System hurtowni danych jest czym znacznie więcej niż prostym narzędziem do raportowania. Jest to środowisko zbierające dane ze wszystkich systemów działających w organizacji. W hurtowni przechowywane są dane historyczne, charakteryzujące się wysoką jakością, zoptymalizowane pod kątem dostarczenia w jak najkrótszym czasie pożądaną informację.

Takie informacje stanowią doskonałe źródło wiedzy na temat działania organizacji, wiedzy, która umożliwia podejmowanie właściwych decyzji oraz zapewnia monitorowanie i kontrolę realizowanych przez organizację funkcji. Dane z hurtowni to także podstawa do tworzenia zbiorczych zestawień statystycznych, które w przypadku działań administracji publicznej są szczególnie istotne zarówno dla samej organizacji jak i szerokiego kręgu odbiorców, do których są adresowane.

Przedsięwzięcie budowy systemu hurtowni danych nie jest proste. Są to zazwyczaj duże i skomplikowane systemy, przy realizacji których występuje dużo zagrożeń. Jednak przy odpowiednim podejściu organizacyjnym oraz doborze właściwych technologii możliwe jest odpowiednie zaplanowanie prac, przewidzenie ewentualnych problemów, zarządzanie ryzykiem i efektywna realizacja założonej funkcjonalności.

W rezultacie włożony wysiłek okazuje się adekwatny w stosunku do korzyści, jakie przyniesie zrealizowany system hurtowni danych, dostarczający informacji niezbędnych na potrzeby własne jak i otoczenia zewnętrznego, oraz które w przyszłości mogą okazać doskonałym źródłem danych na potrzeby opracowywanego przez Komitet Badań Naukowych projektu „e-Government Wrota Polski” zakładającego zintegrowanie usług publicznych świadczonych przez organy administracji publicznej w Polsce.

Projekt „Wrota Polski” zakłada między innymi usprawnienie przepływu informacji do obywatela oraz zapewnienie wymiany informacji pomiędzy urzędami. Zbierający informacje na poziomie całego kraju system może w dalekiej przyszłości stać się doskonałym źródłem dla realizacji ogólnopolskiej hurtowni danych, która zapewni niezbędne informacje pozwalające na podejmowanie trafnych decyzji taktycznych i strategicznych Polski na arenie międzynarodowej.